PM2/APM4 EVALUATION

REPORT

PREPARED BY:
Edwin Wong, PhD, MA
Lingmei Zhou, MS
Joshua M. Liao, MD, MSc

Value & Systems Science Lab
University of Washington
School of Medicine
1959 NE Pacific Street
Seattle, WA 98195

Background. This review document (hereafter, "review") contains feedback from our team at the Value & Systems Science Lab at the University of Washington School of Medicine (hereafter, "VSSL team") pertaining to a retrospective observational analysis of APM4.

The analysis is summarized in a report from Spring 2022 ("Value-Based Payment (PM2)/ Alternative Payment Model 4 (APM4) Final Evaluation.docx"; hereafter, "the report") that consisted of three sections: the first evaluating the relationship between the model and a total costs of care outcome, the second evaluating the relationship between the model and utilization outcomes, and the third evaluating the relationship between the model and quality outcomes. A difference in differences (DID) approach was used in all sections.

This review offers comments on information from the report. Because the report also referenced prior work from Doug Conrad and a team at the University of Washington evaluating APM4 (hereafter, "Conrad team"), this review also notes components from that evaluation as relevant based on documents provided ("UW APM4 Impact Paper.docx" and "HW SIM Eval PM2 Impact Paper (Final to Milbank 8.31.2020) For Editors Eyes Only.docx").

General Comments. We offer several general comments based on our review. The first is in respect to statistical methods. While DID is generally considered among best practices for large-scale policy evaluations, in this particular case, the approach and related methodological decisions involve several limitations. These limitations may influence the accuracy and precision of model parameter estimates of APM4 effects, which should be addressed in any future work.

The second comment pertains to outcome variables. The ability to define appropriate outcomes – and in turn, to draw inference using them – appeared restricted by practical considerations, including data availability and limitations of claims data. Any future work should include refinement of outcome variables to provide insight the effects of APM4. The third comment relates to interpretation of analytic results. Findings should be framed and interpreted carefully in the context of limitations stemming from statistical methods and outcome variables, among other factors.

Our review provides detailed comments on these three issues, as well as others (e.g., theoretical expectations, data sources). These comments are organized using section and subsection headings from the report.

EFFECTS ON 'TOTAL COSTS OF CARE'
Comments by section

Prior work

The report for the final evaluation conducted by the Conrad team is in the process of being published in peer-reviewed literature (Conrad et al. Journal of Healthcare for the Poor and Underserved. 2022. *In press*).

The report notes that in prior work, the Conrad team "was able to detect small, but statistically significant, savings in professional service costs for children assigned participating FQHCs." However, based on documents available to the VSSL team, analyses by the Conrad team appear to identify savings for pharmacy prescriptions among pediatric populations, not professional services (see *Table 4: PM2 Marginal Effects of Payment and Utilization: 18-Month Approach* in both the "UW APM4 Impact Paper.docx" and "HW SIM Eval PM2 Impact Paper (Final to Milbank 8.31.2020) For Editors Eyes Only.docx" files).

Defining 'Total Costs of Care'

We agree with the point being made as we understand it – that payments made for healthcare services do not necessarily reflect the cost of delivering care or health care consumption in dollar terms. However, while there may be unique features of the HCA approach relevant to APM4, this issue – that paid claims do not equate to incurred costs – is not unique to this situation. With that context, the definition of *total cost of care* used (paid claims) is consistent with prior work in the field (see Kaufman et al. Medical Care Research and Review. 2017).

Analysis 1: Member-Months

The primary analysis implements the DID using a mixed fixed effects model with member-months as a unit of analysis. The rationale of the fixed effects approach is to account for unobserved time invariant factors that may confound the relationship between APM4 and outcomes.

Fixed Effects Approach
We recognize the reasons for using fixed effects in this analysis. However, there are two potential tradeoffs in this approach. First, the use of fixed effects is less efficient than alternative models and may yield more conservative (larger) standard errors compared to other candidate approaches. Second, the fixed effects model is susceptible to the incidental parameters problem when the number of observations per member is not sufficiently large. This problem can materialize when a fixed effects model is estimated using a least squares dummy variable approach, as applied in this analysis of APM4. One way to avoid this problem is to estimate models using alternative estimators including the within estimator. Alternatives may also have the advantage of producing more precise coefficient estimates and smaller standard errors.

Assumption of Intention-to-Treat
While we acknowledge the point made regarding potential issues with an "intention to treat "approach, it is important to remember that it is consistent with prior work in payment model evaluation. Additionally, the alternative approach – sometimes referred to as "per protocol" – also has significant limitations worth mentioning.

Member-month Unit of Analysis
With regards to the member-month least squares dependent variable design described:

We agree that an important caveat is that it treats each member month independently. As the evaluation notes, a major limitation is the inability to count for care types that drive spending across multiple months. While the report states that this may be an acceptable compromise, we would note that there are in fact (a) valid reasons to believe that costs can span across months and (b) potential theoretical reasons why high cost acute or chronic events could be concentrated in members based on FQHC APM4 participation status. A member-month modeling approach should use additional measures to account for these issues.

Treatment of Standard Errors

While the report notes that standard errors were clustered at the member level, we believe clustering at the higher level – in this case, FQHC – would be more appropriate. Prior simulation studies have shown that standard errors that are clustered at a level more granular than the level in which variation in the treatment occurs will lead to standard error estimates that are not conservative enough. The implication is that evaluations may identify a treatment effect when one does not exist. (see Bertrand et al. Quarterly Journal of Economics. 2004).

Limitations of Fixed Effects Approach

A preceding point – about potential theoretical reasons why certain factors may vary by APM4 participation status – also extends to a subsequent point made in the section about randomness of attrition. A detailed discussion of theoretical underpinnings aside, it is important to note that all observation analyses are potentially susceptible to differences in time-variant effects such as attrition by assignment group, even with the use of fixed effects.

Specification of Cost Outcomes

It is worthwhile to note the pros and cons of using least squares to model costs and leaving the outcome variable untransformed. The report includes an example about expensive cancer care versus office visits, noting that elimination of several visits represents a linear relationship. We believe it is beneficial to acknowledge that as a (testable) assumption. Using that particular example, it is unclear why even the elimination of several office visits must necessarily reflect a linear relationship for a cost outcome, particularly given likely variation in the underlying utilization patterns of patients receiving care through FQHCs. These limitations are relevant given the descriptive results reported where the mean total costs of care was $333 but the median amount was $0 – highly skewed, as the report describes.

Even in the absence of an *a priori* hypothesis about exponential effects on costs, there are other potential reasons why models with untransformed costs may generate misleading results. With a right skewed distribution in costs, alternative approaches other than the linear model should be considered because high-cost outliers may contribute disproportionately in the estimation of the APM4 treatment effect. Other alternative approaches, including a Tobit model, log transformed one-part model, or multi-part models that account for zero cost observations, can be considered in future work. Finally, it is unclear whether costs were adjusted for inflation, such that all observations represent costs constant for a single year.

Anticipatory Effects

With respect to the issue of anticipatory effects of treatment, we agree that this can be a problem in general for any model or program in which participants have for knowledge of the initiation date. Instead of applying an assumption about whether anticipatory effects are or are not present, or varying the beginning of the intervention period, another approach would be to use a washout period as a sensitivity analysis or robustness check.

Presentation of Results

We believe that showing a broad set of unadjusted results would be beneficial prior to showing adjusted results. Notable unadjusted results include trend plots of key outcomes, tables presenting univariate statistics of outcomes and descriptive comparisons of covariates. The presentation of unadjusted results is important to elucidate notable characteristics of the treatment and control groups, frame the relative scale of estimate effects, and provide transparency in how statistical modeling influences outcomes.

Interpretation of Results

With regards to interpreting results from the regression models, we would offer several perspectives:

- While there is no universal consensus on the alpha level to use to determine significance, we believe it is appropriate to use a 0.1 level in the context of a range of alpha values (0.01, 0.05, 0.1). This is based on prior work and ongoing use in formal evaluations of both primary care and non-primary care payment model evaluations. Consequently, while we acknowledge limitations in using that level (and more broadly, those inherent to using any level of statistical significance), we believe it is too restrictive to a general view of it as not meaningful.

- At the 0.1 level significance, we would underscore the need to interpret the $8/month savings in the appropriate policy and clinical context. The report notes that such savings are ~2% of average spending values over $330. We would point out that this interpretation reflects the notion that incentives from primary care payment models lead to changes in primary care delivery that affect total (primary care + all non-primary care) spending.

An alternative viewpoint is that incentives from primary care payment models lead to changes in primary care delivery that affect primary care savings. To our knowledge, primary care spending was not calculated as an outcome, precluding the ability to interpret $8/month in that context, but any savings would inherently be a larger proportion of primary care versus total spending.

Parallel Trends Assumption

We agree with the comment in the report that by virtue of being an assumption, parallel trends is not directly testable, but that certain tests can provide evidence that may provide more or less support for the assumption. We offer several comments about the testing approach described in the report:

- Though pre-trends analysis is helpful, it does not absolutely confirm or rule out the presence of parallel trends, which is an unobservable counterfactual.
- The pre-trends analysis conducted in the report provides some evidence that the assumption of parallel trends does not hold. For this reason, caution is warranted in interpreting results from analyses conducted using the DID approach.
- Findings from the pre-trends analysis suggest that it may be beneficial to explore other approaches beyond difference-in-differences, such as synthetic controls or other strategies (see Abadie et al. Journal of the American Statistical Association. 2010).

The report notes that because of the pre-trends analysis, the estimate of $8/month should be regarded as a bounded maximum. We believe that is too restrictive a statement, as bounded maximum estimates depend on a number of modeling factors, including those highlighted above.

Analysis 2: Member Years

We appreciate the comment in the report regarding individuals who churned through Medicaid, and the differences between individuals who do and do not churn. However, we do not necessarily interpret the APM4 requirement of 11-12 months of enrollment in a year, as problematic selection.

One alternative view is that APM4 was designed around a subset of patients, not the general served population; and in turn this should encourage providers to enact care changes for that subset. Specifically, care delivery changes APM4 could be expected to benefit individuals enrolled over a prolonged period whereby approaches such as care coordination have enough time take effect. It is unclear why one would expect that a model targeting a subset should necessarily affect the care of the overall population, including those who churn frequently through Medicaid. Independent of other modeling considerations, analyses of APM4-eligible individuals would help generate findings that help address questions about the impact of payment model design and scope.

While the report conveys lower interest in a member-year analysis, we would also highlight the potential appropriateness given that APM4 was targeted to a subset of patients defined by continuous/near continuous enrollment.

Conclusion

We believe that some statements in this section would benefit from qualification. Second, a number of modeling considerations would affect assessment of whether – and to what extent – treatment affects may be overestimated. First, rather than a "known parallel trends violation," it is more appropriate to say that pre-trends analysis does not corroborate the assumption.

SUMMARY

The report details a thoughtful approach to evaluating the relationship between APM4 and total costs of care. Many aspects of modeling assumptions and approach are well-reasoned, and the pre-trends analysis was an important step for assessing the appropriateness of a DID model.

Integrating points made in the report with our perspective and experience, we offer several additional points for consideration:

- All modeling approaches have pros and cons. In considering the objectives of the analysis and study design, there are several methodological alternatives that better capture the data generating process, and yield empirical results that more accurately capture the effect of APM4.

- Based on results of the pre-trends analysis, other methods beyond DID could be explored in the future for evaluation.

- Findings that are statistically significant at the 0.1 level should not be uniformly considered unmeaningful. Beyond statistical significance, clinical and policy significance should be applied when interpreting findings. In this case, such perspectives highlight that other cost outcomes besides total costs of care, such as primary care costs, are worthwhile assessing in future work.

EFFECTS ON ‘UTILIZATION OF CARE’
Comments by section

Introduction

From our perspective, the concept of utilization of care is a multifaceted one, but not a fuzzy concept as reflected in the report. We agree that utilization can encompass many different outcomes, including the three evaluated in the report.

We offer the following perspectives as potential theoretical expectations for the outcomes used in the report:

- Emergency Department (ED) Events

    - There are reasons to expect that in many cases, ED events are undesired both from the patient perspective as well as the health care system perspective. There are many services that can be delivered in either ED or non-ED outpatient settings, and in these cases, it’s reasonable to assume that care in non-ED settings is preferrable.

    - One albeit indirect way to identify such situations is to assess ED events that lead to discharge home/prior location versus events that lead to hospital admission. The expectation would be that the former generally reflect services that could be provided in other non-ED outpatient settings; and the latter generally reflect services that reflect necessary acute care utilization. However, there are exceptions. For instance, chronic conditions that are sub-optimally managed in non-ED outpatient settings may worsen, leading to urgent care needs addressed in the ED and subsequently the hospital – one could interpret these events as circumstances that could be averted with appropriate outpatient care.

    - Ultimately, as with any measure, there is some imprecision associated with assessing ED services as an outcome. Acknowledging this problem, other groups have tried to address this issue by measuring *avoidable* or *preventable* ED visits, or utilization for ambulatory-care sensitive conditions.

    - Regardless, as a utilization measure, ED use is not without theoretical expectations or rationale.

- Primary Care Events

    - With respect to theoretical expectations, an important consideration for primary care utilization is whether (a) events are for visits, procedures or other events; (b) visits are for preventive versus sick care; and (c) visits are related to specific care types. Additional considerations involve identification of relevant patient subgroups (e.g., those who are versus are not up-to-date on guideline-concordant preventive services).

    - In the absence of these distinctions, there are limitations in interpreting findings from analyses using primary care events as an outcome. Conversely, in the presence of these distinctions, a set of theoretical expectations can be applied. For example, to the extent that preventive care is desired under new payment models, one expectation is that preventive primary care events would increase under models such as APM4, at least for those who have historically underused preventive services, and that sick care services would decrease.

- Total Claims

    - In contrast to ED and primary care utilization, we believe that total claims lack detail as an outcome, and likely limited insights to be derived its use as a utilization outcome.

    - For instance, it is hard to know how to interpret the total claims outcome when values range widely (range 0 to >2,000, median of 9 and mean of 26) and there is limited ability to understand if claims reflect visits versus procedures vs medications vs other. We would suggest alternative measures in any future evaluation.

Unit of Analysis, Variable Definition, and Analytic Strategy

The report describes that a member-year approach was used to evaluate utilization in order to capture utilization across a period that is longer than a month ("The year was preferred to the member month for the simple reason that most member-months would be expected to include no primary care utilization, and which months those are for a given recipient is likely random; that is, there is no covariate to give the model that will help it distinguish why someone had a primary care visit in March as opposed to February."). As additional perspective, it is important to note that utilization over time is not completely random. In fact, utilization of primary care, EDs, hospitals, and other sites of care are often unevenly distributed across time. This may be due to regularity in chronic care needs and/or deterministically time-clustered acute care episode needs. For instance, visits to urgent care or ED settings can be clustered with hospitalization. As another example, individuals can be particularly vulnerable of hospital readmission in the time period immediately following a preceding hospital admission.

In turn, we agree with the member-year approach used. However, we would note that the underlying rationale here raises a question about the modeling approach used in the 'total cost of care' section, which reflects a preference for a member-month analysis.

With respect to variables:

- There are multiple ways to define ED events and primary care utilization, which include the approaches used to define "events" in the report, but also others beyond them.

- For primary care, it is worth noting that the narrow definition used in the OFM 2019 report has certain potential problems. In particular, taxonomies used in the narrow definition may need adjustment in order to more appropriately include primary care, and exclude non-primary care, services.

As APM4 is targeted to FQHCs, which represent important sources of primary care, these potential problems are salient.

- Based on our understanding about CDPS and its use in associating clinical severity with future costs, inclusion of that variable appears appropriate for total cost of care outcome above. However, it is less clear why CDPS, as opposed to other measures of clinical severity, would be the best variable to include in analyses of utilization outcomes. While one perspective is that utilization trends with costs, it is also relevant to acknowledge that utilization patterns and utilization of different types of care lead to variation in that relationship.

- As we understand it, utilization analyses included fixed effects for member characteristics such as race, gender, and ethnicity, but not member fixed effects, as were used in cost analyses. It would be beneficial to consider, or describe limitations in using, member fixed effects in utilization models.

Analytic Strategy

The analysis operationalizes the DID approach to analyzing binary outcomes using a logistic regression with dummy variables capturing member fixed effects. Similar to analyses of costs, there is risk of incidental parameters bias, where the effect of APM4 may be incorrectly estimated.

The clustering of standard errors at the member level is also expected to underestimate the true uncertainty of parameter estimates, resulting in the risk of over-rejecting the null hypothesis that APM4 has no effect.

The report indicates a pre-trends analysis was not possible with member-year data. However, it is unclear why such an analysis was not possible if there are multiple years of data prior to the implementation of APM4. In the absence of absolute limitations, we believe that conducting a pre-trends analysis is important to the analytic strategy.

Analyses of utilization involved logistic regression models. This has the limitation of informing the effect of APM4 on any use, but not intensity of use. Further analyses should consider the use of count data models, particularly for utilization outcomes with variation in its distribution.

Results

We believe that showing results as probabilities would support interpretation.

We believe that showing unadjusted results for all outcomes (these were reported for ED Events and Total Claims, but not Primary Care Events) would be beneficial prior to showing adjusted results.

Conclusion

As noted above, we believe there are theoretical expectations for utilization outcomes. It is not clear why analysis of utilization would be necessarily viewed as exploratory.

We do believe that results should be caveated, and potential future evaluation work should be encouraged, in view of the issues noted in prior sections.

SUMMARY

The report details a clearly described approach to evaluating the relationship between APM4 and utilization outcomes. Integrating points made in the report with our perspective and experience, we offer several additional points for consideration:

- Other outcomes – both variations of those assessed, and others altogether – should be considered in future evaluation. Outcomes can be identified and prioritized based on clinical and theoretical expectations for utilization changes under payment model incentives.

- Similar to analysis of cost outcomes, analysis of utilization outcomes are subject to tradeoffs in analytical modeling choices. We believe in this case, there are methodological alternatives that are more appropriate given the structure of the analytical dataset.

EFFECTS ON ‘QUALITY OF CARE’
Comments by section

Introduction

We agree it is instructive that “According to program staff, during the duration of the PM2/APM4 program, until the outbreak of the COVID-19 pandemic, no participating FQHC had failed to achieve levels on these quality measures that entitled the FQHC to full payment.” While APM4 participants could still improve on these metrics, incentives to improve generally tend to be weaker when one is already above a performance threshold versus below it.

We would also point out the apparent discrepancy between the way APM4 metrics are listed in the report, versus how metrics are listed in the Conrad team evaluation documents (“UW APM4 Impact Paper.docx” and “HW SIM Eval PM2 Impact Paper (Final to Milbank 8.31.2020) For Editors Eyes Only.docx” files).

The report lists 9 metrics:

1. *Comprehensive diabetes care - poor HbA1c control (>9%)*
2. *Comprehensive diabetes care - blood pressure control (<140/90)*
3. *Controlling high blood pressure (<140/90)*
4. *Antidepressant medication management: Effective acute phase treatment*
5. *Antidepressant medication management: Effective continuation phase treatment (6 months)*
6. *Childhood immunization status - combo 10*
7. *Well-child visits in the 3rd, 4th, 5th and 6th years of life*
8. *Medication management for people with asthma: medication compliance 50% (Ages 5-11)*
9. *Medication management for people with asthma: medication compliance 50% (Ages 12-18)*

In contrast, one document from the Conrad team (“UW APM4 Impact Paper.docx”) describes 7 APM4 quality metrics, with the following detail:

> *"Three of the seven metrics are outcome (not process) measures for diabetes care; two are for medication management (for anti-depressants and asthma medicines, respectively); another for childhood immunization status; and one for well-child visits."*

The other document from the Conrad team ("HW SIM Eval PM2 Impact Paper (Final to Milbank 8.31.2020) For Editors Eyes Only.docx") appears to identify 8 measures with the following detail:

> *"Three of the eight metrics are outcome (not process) measures for diabetes care; three are for medication management (for anti-depressants in the acute and continuous phases, respectively, and asthma medicines); another for childhood immunization status; and one for well-child visits."*

This discrepancy appears to be one of description, not content: based on Appendix information included in the Conrad team documents (in Appendix Table 1 in both, there are 7 metrics listed in both files, with two parts – a and b – for metrics 4 and 7), both include the same measures listed in the report. However, it may still be important to understand how metrics are counted (as 7, 8, or 9) to the extent that methods for calculations vary based on counting method. For instance, performance on "antidepressant medication management" may vary based on whether that is one metric based on performance on both acute phase and continuation phase treatments; versus if acute phase treatment and continuation phase treatments are separately calculated as two different metrics. It is important to have clarity on how metrics are referred to and calculated across documents particularly if there is a desire to make comparisons.

Prior Work

To our understanding from the report, there were four sets of quality metrics being discussed in the report:

> Set A. The set of 9 metrics included in APM4 (and enumerated in the "Introduction" section), available as part of required reporting. One benefit of assessing these metrics is that they are available across the study period,

and theoretically, incentives would be stronger given inclusion in APM4. One limitation is that these metrics were only available for participating FQHCs, but not non-participating FQHCs, limiting performance comparisons. Data collection represents another area of potential limitation (need to self-report, validate data; potential for missingness).

Set B. The set of 9 metrics included in APM4, available via MCO-submitted MLD. Assessing these metrics has the same benefits and limitations as Set A, with the exception that data were available only for 2016 and following, but were available for both participating and non-participating FQHCs. Another limitation is that these were sampled rather than obtained from the total population.

Set C. A subset of the 9 metrics included in APM4 (not enumerated in the report) and available as part of HCA-produced metrics outside of required APM4 reporting. One benefit of assessing these metrics is that they are not subject to the same data collection issues. One limitation is that they were only available starting in 2017, and therefore could not be assessed prior to initiation of APM4.

Set D. Described as part of prior work from the Conrad team, this is a set of common ambulatory care metrics not directly included in APM4. Of the metrics reported in the "UW APM4 Impact Paper.docx" and "HW SIM Eval PM2 Impact Paper (Final to Milbank 8.31.2020) For Editors Eyes Only.docx" documents, 1 metric appears to correspond in focus to the set of 9 metrics above (metric for A1c testing in patients with diabetes). One benefit of assessing these metrics is that performance can be compared between participating versus non-participating FQHCs. One limitation is that these metrics were not included in APM4.

We agree with the report that two major issues in evaluation are (a) the absence of data in the pre-intervention period and (b) the absence of data for a comparison group. Set A suffers from the absence of comparison group data, while Sets B and C suffer from the absence of pre-intervention period data. Unfortunately, we believe the inadequacies in Sets A-C are foundational, and that it is not possible to achieve adequate results (estimates of the impact of APM4 on quality) by analyzing inadequate data. Analyses would need datasets that overcome these limitations.

To that end, we believe that Set D should be considered for evaluation. Acknowledging the value of assessing metrics that are included in APM4, this benefit is partially offset in the case of APM4 by a point made in the "Introduction" section – that all participants were already meeting performance thresholds for all of the 9 metrics. As we understand it, the value of Set D appears to be the ability to conduct evaluation using pre-intervention and intervention period data, for participating and comparison groups.

Dataset Construction and Analytic Strategy

This section of the report notes that analysis was conducted on a population subset that had 11 or 12 months of enrollment in a given year, consistent with APM4 rules. Please see our comments above in the "Effects on 'Total Costs of Care'" section regarding use of this population for analysis.

Analyses did not appear to include adjustment for clinical severity using measures such as the Charlson Comorbidity or Elixhauser indices or conditions. While each measure possesses pros and cons, standard practice would be to include some form of risk adjustment in the analysis.

Similar to the analyses of cost and utilization, the use of fixed effects and clustering of standard errors at the member level creates challenges and potential inaccurate estimates of the effect of APM4 on quality. Please see our comments in preceding sections for more detail.

Results

Please see our comments above in the "Effects on 'Total Costs of Care'" section regarding the use of alpha levels to determine statistical significance, and the need for clinical and policy perspectives in interpretation. These issues apply to how we would interpret the findings from this model.

Analytic Strategy 2: The Within Model

We appreciate the thoughtfulness expressed in the report, weighing pros and cons of different datasets. As noted above, however, we believe that when the evaluation goal is to understand the treatment effect, the greatest threat to validity arises from issues that confound treatment effect with other effects. In this case, that pertains to Set A (data submitted as part of required reporting and available only for participating FQHCs across the study period).

Results

We believe it is prudent to avoid over interpretation of trend plots, as suggested in statements about obvious impacts from visual examination of plots.

We believe that showing unadjusted results (either in text or tables) would be beneficial prior to showing adjusted results.

Conclusion

We agree that taken together, limitations to different sets of data preclude optimal evaluation. As noted above, we believe that certain limitations are greater threats to evaluation efforts than others, suggesting that prioritizing analyses and datasets based on severity of threat, and conducting multiple analyses using multiple datasets, would be prudent measures for evaluating quality metrics.

SUMMARY

The report details a thoughtful approach to evaluating the relationship between APM4 and quality of care. The thinking behind modeling assumptions and approach are clearly articulated. Integrating points made in the report with our perspective and experience, we offer several additional points for consideration:

- Given the major challenges using Sets A-C to evaluate changes over time or by participation status, future evaluation should consider evaluation using Set D or additional outcomes that permit the use of a comparison group and pre/post periods.

- Analysis using Set D (common ambulatory care metrics not directly included in APM4) or other outcomes can be justified in the context of pros and cons of different data sources, and the potential ability to provide comparisons to prior work and assess the appropriateness of a DID method.

- Adjustments to the described model parameters or approach could be beneficial.

- Findings should be interpreted not just through the lens of statistical significance, but clinical and policy significance. That is particularly relevant for this analysis, given the benefit of analyses on multiple metrics using multiple datasets.